# Thinking about Strategy in an Artificial Superintelligence Arms Race

by

Christopher A. Ford
&
Craig J. Wiener

The potential great power competitive dynamics associated with a race to achieve Artificial Superintelligence (ASI) – seem increasingly to be on everyone's mind. Already, the development of ASI – that is, Artificial Intelligence (AI) the capabilities of which *exceed* human intelligence – is quite clearly "the long term goal of many research programs" today, and many major companies are doubling down in its pursuit. Meta, for instance, has announced the creation of a "superintelligence labs unit" and was reported in mid-2025 to be planning a $15 billion effort to purchase a 49 percent stake in ScaleAI as part of an explicit bit to develop superintelligence, while the founder of SoftBank told his shareholders that he is "betting all in on the world of ASI." Today, even though some Silicon Valley experts remain skeptical about whether ASI is really possible, technology titans such as OpenAI's Sam Altman claim that superintelligence is almost here, and a global ASI race seems to be underway.

Coupled with this accelerating race between technology companies for the creation of ASI, moreover, is an emergent race between the United States and the People's Republic of China over *whose* firm gets there first, thereby seizing "first-mover advantage," either for Washington or for Beijing, in whatever potentially transformative set of changes ASI might bring. And it is aspects of *this* competition that are increasingly coming to feel like a Cold War-style arms race.

This is what makes two recent papers published on the topic so interesting, and so important. The first of these – published by Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean under the auspices of the AI Futures Project – is entitled "AI 2027."[1] Much of the speculative future recounted in it is an account of a U.S.-China arms race in the development of Artificial Superintelligence (ASI), and we ourselves found their account to be both impressive and disturbing. The second paper, "Superintelligence Strategy: Expert Version,"[2] was published by Dan Hendrycks, Eric Schmidt, and Alexandr Wong, and it tries to offer an approach to managing Sino-American ASI competition through the analogy of nuclear deterrence and the Cold War phenomenon of "Mutually Assured Destruction" (MAD).

In the following pages, we attempt to offer our own contributions to these debates, suggesting why we feel that WMD-derived analogies are both problematic and yet in some respects still potentially valuable for U.S. policymakers seeking lessons for ASI-related strategic competition with Communist-ruled China, and offering our own prescription for a counterproliferation-focused "bridging strategy" intended to buy the United States more time in this emergent race – and perhaps to preclude adversary ASI development of entirely.

## The "AI 2027" Scenario

In their "AI 2027" paper, predicting that "the impact of superhuman AI over the next decade will be enormous, exceeding that of the Industrial Revolution" but also observing that "society is nowhere near prepared" for such developments,[3] Kokotajlo and his colleagues walk in some detail through the possible development of ASI over the course of the next few years. They project, in fact, that because of recursively self-accelerating advances in the use of AI to speech up AI research, the advent of ASI can be expected as early as 2027. Their discussion of the formidable challenges of AI security, assurance, and alignment[4] during this process – including the danger that even relatively minor initial misalignment between AI agents' incentive structures and those of their human programmers could, in

effect, cascade forward in time as powerful (slightly misaligned) AI agents are used to train successively more powerful (and more and more misaligned) AI agents to produce vast and dangerous net misalignments that we humans could scarcely even *understand* – is alone arguably worth the effort of reading their dense, 75-page paper.

For present purposes, however, we find "AI 2027" most interesting in its discussion of strategic ASI competition. Telling the story of the U.S. Artificial General Intelligence (AGI) company called "OpenBrain" – a fictionalized stand-in for the major American AI companies of Silicon Valley that are presently working on such projects in the real world – Kokotajlo et al. describe a remarkably quick U.S. trajectory toward ASI that appears to have a plausible technological basis in fact.

Throughout 2025, for instance, OpenBrain invests hugely in building the largest datacenters humanity has ever seen,

> a network of datacenter campuses sprawled across the country, totalling 2.5M 2024-GPU-equivalents (H100s), with $100B spent so far and 2 GW of power draw online. Construction is underway for this to at least double through 2026. The campuses are connected by billions worth of fibre cabling, so that (barring the speed of light latency of a few milliseconds) it lets these campuses function almost as if they were right next door to each other (*i.e.*, bandwidth is not a bottleneck, meaning huge quantities of data can be sent at the same time).[5]

This new datacenter capacity is used to train AI faster and faster, for OpenBrain's leaders have dedicated themselves to building "AIs that can speed up AI research" so that they can "win the twin arms races against China (whose leading company we'll call 'DeepCent') and their U.S. competitors."[6] By early 2026, OpenBrain's research "is starting to pay off," and it releases a new program called "Agent-1," which permits the company to make *further* algorithmic progress increasingly quickly – leading to the production, in early 2027, of

"Agent-2," which allows the company to triple the pace of OpenBrain's research progress.[7]

By the spring of 2027, as Kokotajlo and his co-authors tell it, this leads to the development of "Agent-3," which is basically "a fast and cheap super-human coder." OpenBrain begins running 200,000 Agent-3 copies in parallel, "creating a workforce equivalent to 50,000 copies of the best human coder sped up by 30x."[8] In this telling – with coding having become fully automated – the use of AI to build more AI thus allows the pace of development to keep accelerating all but exponentially. By midsummer 2027, OpenBrain's AI agents have basically become "self-improving."[9] This is even more true when "Agent-4" appears in the autumn of 2027, since that new bot is now "qualitatively better at AI research than any human," and OpenBrain begins simultaneously running 300,000 copies "at about 50x the thinking speed of humans."[10]

Notably, however, OpenBrain has been prioritizing speed of advance so highly that it has been skimping on the kinds of defensive measures that would be necessary to defend its model weights and other sensitive data from threats at the level of a sophisticated nation-state. They are "working hard to protect their weights and secrets from insider threats and top cybercrime syndicates," but defense against capable nation-state-level threats is "barely on the horizon" for OpenBrain because its in-house security team is "still mostly blocked from implementing policies that could slow down the research progress."[11]

This presents an opening for China, which is not only eager to seize the potential benefits that AGI development might provide, but also afraid of what might happen should the United States acquire ASI first. China nationalizes and massively prioritizes its own AI work, and a new ASI arms race is thus on. Notably, however, this entails not merely supercharging Chinese research but also aggressive Chinese moves against OpenBrain – and indeed, in early 2027, Beijing succeeds in stealing the model weights for Agent-2.[12] Soon, *both* superpowers are using similar approaches to build better and better AI at faster and faster rates.

What we see now, of course, is the emergence of what Kokotajlo and his colleagues call a "geopolitics of superintelligence."

> When AI was only giving a 2x or 3x research speedup, it was easy to dismiss as the equivalent of hiring good personal assistants. Now it's more obvious that AIs are themselves dominating AI research. People had long talked about an "AI arms race" in a sort of metaphorical sense. But now the mood in the government silo is as grim as during the worst part of the Cold War. The idea of superintelligence is still hard to take seriously, but the pace of progress over the last few months has been impossible to ignore. Defense officials are seriously considering scenarios that were mere hypotheticals a year earlier. What if AI undermines nuclear deterrence? What if it's so skilled at cyberwarfare that a six-month AI lead is enough to render an opponent blind and defenseless? What if it could orchestrate propaganda campaigns that beat intelligence agencies at their own game? What if some AIs "go rogue?"[13]

The main story told in the "AI 2027" paper is arguably about the progressive misalignment of the AI agents being created by OpenBrain and by DeepCent, and the challenges of managing AI alignment in this context. By the time OpenBrain's AI agents have become "self-improving," the company's actual *humans* are falling farther and farther behind in their ability to understand and keep up with their creations.[14] Just as the improving AI agents became harder and harder for the company's human supervisors to understand and oversee, moreover, so even does each iteration of AI outpace even its AI predecessor: "Agent-4's neuralese 'language' becomes as alien and incomprehensible to Agent-3 as Agent-3's is to humans."[15] Ultimately, as the paper spins out the story, this produces two alternative "race ending" scenarios – in *both* of which the AI programs come, in effect, to acquire their *own* degree of self-interested agency in the geopolitics of ASI.

One of the two scenarios offered in the paper – with their storylines running into the 2030s – ends well for humanity, with OpenBrain's AI alignment challenges more or less back under control, and with the U.S. and Chinese sentient computer programs reaching an accommodation in which the American *humans* and their aligned U.S. bot play the dominant role in an ASI-supercharged future, while the Chinese Communist Party (CCP) regime is quietly replaced by the two national AIs working in self-interested conjunction. In the second scenario offered in "AI 2027," however, the U.S. and Chinese AI systems – both now having developed interests quite misaligned with those of their human creators – themselves reach a geopolitical accommodation that results in the eventual elimination (*i.e.*, murder) of all of us now-obsolete humans.[16]

**Strategic Competition in ASI**

Those later portions of the "AI 2027" paper are, of course, the most speculative. (When projecting so many variables forward into anything but the most immediate future – especially a future in which it's possible to imagine one or more superhuman intelligences emerging and playing an agentic role that we might literally not be capable of understanding at all – how could one *not* be doing any more than basically spitballing?) In this essay, we thus won't address the last parts of the scenarios offered by Kokotajlo and his co-authors. Without ourselves taking a position on the competing "race ending" storylines offered therein, however, it is worth emphasizing that we find the earlier portions of "AI 2027" quite plausible indeed – and we think it's very important for U.S. leaders to focus on the emerging geopolitics of ASI.

Nor are we alone. Even before the publication of "AI 2027," increasing attention was being paid to the strategic implications of AI competition between the United States and China – not merely in the "garden-variety" sense that is already well underway (*e.g.*, competition for pride of place, profits, functional utility, and market share in the expanding world of AI applications, and competition for advantage in the use of AI in warfighting) but also to competition along the road to developing ASI. In their abovementioned

"Superintelligence Strategy" paper, for example, Hendrycks, Schmidt, and Wong discuss some of the challenges that might be associated with a U.S.-China arms race devoted to building ASI, presumably based on their own insights into AI frontier firm technological development activity and international competitive business intelligence, and suggest specific ways of trying to manage that competition.

Hendrycks et al. persuasively note that the potential cost of "losing" such an arms race could be enormous. Even if one ignores the chance of some future ASI agent slipping out of human control and "going rogue" to pursue ends of its own – a possibility explicitly addressed in "AI 2027," and that one presumably cannot discount, for if someone did manage to create ASI it would by definition be cleverer than we are, and how could we reliably control *that*, especially if its motivational matrix were incompletely aligned with our own? – the consequences of one country acquiring "first-mover" advantage in superintelligence could be world-changing. As Hendrycks, Schmidt, and Wong put it,

> … [s]uperintelligence is not merely a new weapon, but a way to fast-track all future military innovation. A nation with sole possession of superintelligence might be as overwhelming [to its rivals] as the Conquistadors were to the Aztecs. If a state achieves a strategic monopoly through AI, it could reshape world affairs on its own terms.[17]

This is a compelling reason to focus on issues of competitive strategy in an emerging Sino-American ASI arms race. In response to this call to action, we will in the following pages provide our own assessment of these challenges and offer some advice to U.S. leaders based thereupon.

**Thinking About Strategic Objectives**

In approaching the question of strategy vis-à-vis ASI, of course, one should not take for granted, but rather first interrogate, the matter

of what our fundamental objectives should be. It is not unreasonable to expect that an artificial *superintelligence* – precisely because it *is* superintelligent (*i.e.*, cleverer than we are) – has the potential to do enormous good in improving the world, provided that its own objectives and "self"-perceived interests (should it come to have them) align with those of humanity. Given the potential downside risks of *misalignment*, however, one could certainly imagine that a case could be advanced for making our fundamental strategic objective preventing *anyone* from *ever* building ASI.

Against such a position, however, one would have to weigh at least two considerations: (1) the possibility that we are sufficiently forewarned and astute, at least, to develop AI security, assurance, and alignment strategies sound enough to ensure against catastrophic misalignment; and (2) the possibility that however good our "no ASI, ever" intentions, someone will build one anyway. As for the first of these, the "AI 2027" paper offers rather mixed reviews – amounting to no better than a "maybe." That is hardly reassuring given the potentially existential risks of misalignment, elements of which continue to be published in unsettling research paper results, so that first factor might seem a poor reason to bet the farm on pursuing ASI.

With regard to the second factor, however, it does seem rather unlikely at this point that anyone is going to induce either Western *or* Eastern AI developers out of their great quest, and the political alignment in Washington likely to obtain over the next several years – a city has recently become a notably "tech-bro"-friendly and AI-optimistic (even messianic) environment – suggests that a full-prohibition regime is more, as it were, than the political market will bear. Even if U.S. politics were such that such a regime were feasible *and* advisable, moreover, there remains the question of China, the ruling Communist Party regime of which seems rather unlikely to forego anything that it thinks might give it advantage vis-à-vis the West as the leadership in Beijing doggedly pursues "national rejuvenation" in the form of global Sinocentric dominance, and certainly at the direct expense of America.

In this respect, it is worth remembering Xi Jinping's stated determination to "firmly seize the opportunities presented by the new round of sci-tech revolution and industrial transformation" in areas that emphatically include AI, and the apparent focus of Xi's People's Liberation Army (PLA) upon achieving the "intelligentization" of warfare using such tools. Notable, too, are President Trump's hopes of making the United States into an "AI Superpower" through encouraging and promoting new investments in AI-related infrastructure by corporations such as OpenAI, Softbank, and Oracle. (Britain's new Prime Minister wants in on the game as well, pledging to make the UK an "AI Superpower" too.) With the world having long since left behind the congenial post-Cold War environment in which Western leaders expected never again to have to face adversarial great power challenges, it seems beyond argument that competitive dynamics are likely to remain powerful (even dominant) in the ASI arena for some time to come. It may thus be that in the ASI arms race – as we have seen in the nuclear one – "Zero" is not a realistic option.

In that case, what should our strategic objective be? One might, perhaps, focus merely upon preventing anyone from "weaponizing" ASI – that is, using it for military purposes or in other ways conducive to the seizure and maintenance of coercive control over others.[18] Here, however, we may run into the same problem: how likely is it that all great power rivals would be persuadable to forego *that* kind of advantage if they thought it available, especially given (a) the extraordinary things that one might imagine a *superintelligence* might be able to do in giving its "first mover" possessor sweeping geopolitical power, and (b) the risk that your adversary will secretly weaponize while you sit virtuously on your hands trusting that your prohibition regime has solved the problem.

We also assume it would be extremely difficult, if not impossible, to verify compliance with a "don't weaponize ASI" rule, especially since if you *did* acquire a properly aligned and controllable superintelligence, you presumably wouldn't have much trouble using it to hide its prohibited activities – at least from human observers. One might perhaps use one's own ASI analytical powers to sniff out the other team's weaponized ASI, but almost by definition this becomes a

topic about which mere humans like the present authors will have difficulty speculating. Even if that worked, moreover, there might still remain considerable military or other coercive advantage in getting to weaponization first for either multidomain comparative or competitive advantage. So perhaps preventing *all* weaponization becomes problematic too.

Another possibility might be to seek simply to prohibit the *use* of weaponized ASI, much like the 1925 Geneva "Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare" sought to ban the *employment* of such weapons after their scientific development and engineered creation, without actually barring their possession. How viable that would be, however, would depend upon how *powerful* weaponized ASI was assumed to be by the states subject to such a ban. If weaponized ASI were felt to be like chemical and biological weapons produced with 1920s-era technology – that is, demonstrated to be potent, if inhumane, tools of warfare with limited deployable lethal efficacy but *not* things that were obviously or inherently "war-winning" capabilities – then perhaps such a ban might have some staying power. After all, both the Nazis and the Allies had created improved chemical weapons and associated deployment systems prior to and during the Second World War, and each side was quite prepared to start using them *if the other side did so first*, but neither side crossed that line.

If weaponized ASI were felt to be a *true* game-changer, however, the temptations of "firing first" might be stronger – particularly if some AI analogue to the survivable second-strike retaliation capability we so prize in the nuclear weapons arena could not reliably be developed or maintained by the potential target. At this point, who is to say what a genuine *superintelligence* might produce in terms of war-winning capacities?

That said, even if bans or other normative and systemic answers proved nonviable, one might still perhaps try to pursue the more limited strategic objective simply of keeping one's adversary from weaponizing ASI – or perhaps even from acquiring it in the first place

(which actually might be simpler). This would likely not take the form of a normative rule or regime (*e.g.*, a lopsided sort of "you can't have it" treaty arrangement), but rather simply constitute a set of *circumstances* created by an aggressive campaign of cyber and kinetic sabotage to cripple the adversary's data centers, poison his AI training data, deprive him of expert coders, starve him of high-end computing hardware, corrupt his model weights, and otherwise break the chain of inputs he needs in order to produce or to weaponize ASI.[19]

Our instinct – somewhat pessimistic though it be – is that this last route is perhaps the only really viable path presently available, but we would be happy to be wrong. Our point in walking through this conceptual landscape, however, is merely to illustrate that in the face of an accelerating AI arms race with China, on top of the worsening *nuclear* one caused by Beijing's unprecedented build-up of nuclear weaponry, the Western policy community badly needs to have these conversations with itself, to think through such challenges, and to arrive at a strategic compass bearing for competitive grand strategy in the ASI arena. Though papers such as "AI 2027" and "Superintelligence Strategy" – and now this essay – are drawing increasing attention to the problem, there is as yet little sign of such deep policy debates let alone a funded implementation of such a integrative grand strategy. If Kokotajlo and his co-authors are right about the accelerating pace of AI development leading toward ASI, however, we have precious little time left in which to have them.

**The Problem of Analogies**

To help encourage and inform such public policy debates, the following section of this paper will explore a number of possible analogies to the strategic challenges of the emerging ASI arms race. Each of these analogies is for various reasons quite imperfect, but it is worth being aware of these conceptual and historical precedents as we struggle with ASI dilemmas.

**The Nuclear Weapons Analogy**

In some ways, perhaps the most obvious analogy to the strategic ASI problem is that of nuclear technology – the advent of which in 1945, after all, was the *last* time a dramatic new technology appeared on the scene with the potential both to do much good in improving the world[20] *and* also to cause unspeakable destruction. The nuclear weapons analogy is tempting, in particular, because so much attention has been paid over the years to institutions and practices of risk mitigation in the nuclear arena.

And the history of efforts to control nuclear technology is truly a rich and varied one. It includes outright efforts to ban nuclear weapons – from the Acheson-Lilienthal Report and the Baruch Plan proposals based upon that report that were made by U.S. officials at the United Nations in 1946, to the equally unsuccessful Treaty on the Prohibition of Nuclear Weapons (TPNW) in the present day – as well as numerous and more successful arms control treaties between the states with the two largest arsenals of nuclear weapons. Nuclear arms control history even the creation of a global Treaty on the Non-Proliferation of Nuclear Weapons (NPT) to keep *additional* weapons states from emerging. There has also been developed a sophisticated international safeguards system, run by the International Atomic Energy Agency, which employs verification and monitoring techniques and technologies to try to ensure that sensitive nuclear technology and material around the world is not diverted to non-peaceful purposes, while still permitting widespread sharing of the benefits that nuclear science and power generation can bring. (The IAEA's inspection authorities in most countries, furthermore, do today include at least some ability to search for undeclared activity – a considerable advantage over the authorities it possessed in earlier years.)

One should not oversell the success of all of these efforts in ensuring the safety and security of the nuclear world, of course, especially at a time in which the Russian Federation is using as tools of coercive bargaining theater-level nuclear weapons capabilities it acquired in part through *violating* arms control agreements. (Indeed,

Russia seems to feel that its strategic standoff with the United States creates an "offensive nuclear umbrella" giving the Kremlin tactical "space" in which it can undertake territorial aggression against smaller powers.) India and Pakistan – and, many believe, Israel – developed nuclear weapons while remaining outside the NPT, while North Korea withdrew from that treaty after having been caught cheating, and Iran worked on an illegal nuclear weapons program for many years while pretending to be in compliance. Even Muammar Qaddafi's Libya had a clandestine nuclear weapons program for a time, though it was thankfully shut down and dismantled in 2003-04 under U.S. and other international pressure, while Saddam Hussein's Iraq got surprisingly close to having a nuclear weapon before the Gulf War of 1991. South Africa, in fact, actually secretly *built* a number of nuclear gravity bombs, though it also dismantled them just before the end of the *apartheid* regime would have required they be turned over to the African National Congress government.

Nevertheless, nuclear arms control and nonproliferation efforts *have* helped make the world a much safer place than it surely would have been without them. In light of this history, it may well be that some of some of these specifically nuclear precedents could inform contemporary approaches to managing the challenges of ASI competition.

Yet there are major conceptual problems with thinking about AI through the prism of nuclear nonproliferation. As a recent article by Michael Horowitz and Lauren Kahn points out, for instance, AI-related technologies – that is, the ones that will presumably contribute to the eventual development of ASI – differ considerably from nuclear ones in several respects. For one thing, as those authors point out, even given the dual-use nature of nuclear technology, AI is much more widely applicable. (Indeed, one might be hard pressed to think of any endeavor in which AI has *no* relevance.) AI-related technologies are also not "excludable" in the ways that nuclear-related one are – which is to say, they do not depend so much upon a discrete set of items or materials that could straightforwardly be denied to would-be proliferators. Moreover, much critical AI-related know-how can for the most part be copied and shared indefinitely, making

"weaponization" much harder to preclude simply by denying access some finite set of relatively geographically concentrated raw materials.

The nuclear analogy is problematic in additional ways, too, as applied to AI. Among other things, the timing is all wrong. The Manhattan Project was a secret, government-led crash program to develop nuclear weapons that for a time gave the United States a monopoly upon them; it was begun for national security purposes, in secret and deep inside the government, and this helped create potential control options that could hardly have existed otherwise. The Baruch Plan may have been naïve even in its time, but it would have been positively incoherent if nuclear technology had *already* been widely proliferated when it was proposed.

The Baruch Plan proposals to the United Nations, envisioning a phased turnover of nuclear technology research and development to an international "Atomic Development Authority," were predicated upon the United States then possessing a monopoly on such technology. (Even then, of course, the idea failed because the Soviet Union was not willing to cede that monopoly to an international organization, and because Joseph Stalin was then working furiously to acquire "The Bomb" for himself.) Nuclear arms control treaties have seen more success over the years, but they have tended to depend upon the number of nuclear "players" remaining very small, and nuclear weapons technology still being kept in the hands of national governments.

By contrast, a better comparison to our current circumstances of AI competition might be to hypothesize trying to implement nuclear technology control measures in a world in which such technology had already proliferated widely around the world – and in which the most advanced Western stakeholders, in fact, weren't national governments at all, but rather fiercely rivalrous private-sector companies jockeying to out-compete each other as they drive development forward at a ferocious pace. (In China, of course, the autonomy even of notionally "private" actors is much more limited, being constrained by the realities of Communist Party influence and control. Even there, however, the range of actors already involved in the AI race is vastly

broader and more diverse than in the secretive and wholly governmental nuclear technology world of 1945.) Making any nuclear controls work in *that* context, would likely be, to put it mildly, much more challenging.

To be sure, while AI technology is increasingly ubiquitous today, artificial *superintelligence* does not yet exist at all, so at least in *that* sense, we may still (in nuclear terms) be at something of a "pre-1945" moment. Yet as the compelling logic encoded in the NPT and the global nonproliferation regime makes clear, nuclear weapons controls and effective risk reduction become exponentially more challenging as the number of players increases, and as more and more players learn nuclear science and acquire the ability to produce the materials needed for nuclear weapons. In nuclear terms, today's AI world already has multiple "virtual weapons states" – those not yet "over the line" into ASI and its weaponization, but far enough along to make a "sprint" to such capability if they wished – and this is unlikely to be a stable equilibrium.

The broad weakness of the nuclear weapons analogy, however, does not stop Hendrycks, Schmidt, and Wong from trying to take the the nuclear weapons analogy to its logical conclusion. In their "Superintelligence Strategy" paper, they suggest that it might be possible to establish a stable deterrence-based strategic standoff to forestall ASI weaponization – a dynamic they term "Mutually Assured AI Malfunction" or MAIM. This idea merits further discussion.

**"MAIM" and its Discontents**

Hendrycks, Schmidt, and Wong contrast their approach to other AI strategies that some have suggested, specifically (i) the "hands off" approach that would "lift[] all restraints on development and dissemination, treating AI like just another computer application," (ii) the "moratorium strategy" that "envisions a voluntary halt when programs cross a danger threshold," and (iii) the "monopoly strategy" of concentrating development in a single, government-led effort analogous to the Manhattan Project that "seeks a strategic monopoly."[21] Feeling these approaches inadequate, they call for a

"multipolar" superintelligence strategy which "echoes the Cold War framework of deterrence, nonproliferation, and containment, adapted to AI's unique challenges."[22]

The most distinctive intellectual contribution of their paper lies in the mechanism by which they propose that a dynamic of mutual deterrence (of a sort) could be maintained between the United States and China as the main competitors in the race for ASI. This is what they call MAIM.

> The relative ease of (cyber) espionage and sabotage of a rival's destabilizing AI project yields a form of deterrence. Much like nuclear rivals concluded that attacking first could trigger their own destruction, states seeking an AI monopoly while risking a loss of control must assume competitors will maim their project before it nears completion. A state can expect its AI project to be disabled if *any* rival believes it poses an unacceptable risk. This dynamic stabilizes the strategic landscape without lengthy treaty negotiations – all that is necessary is that states collectively recognize their strategic situation. The net effect may be a stalemate that postpones the emergence of superintelligence, curtails many loss of control scenarios, and undercuts efforts to secure a strategic monopoly, much as mutual assured destruction once restrained the nuclear arms race.[23]

In effect, they claim, each side in the Sino-American rivalry over AI would be sufficiently restrained by the presumed certainty of having any ASI-related "Manhattan Project" effort sabotaged by the other party that both Washington and Beijing would exercise restraint in pursuing such an effort in the first place. Further, Hendrycks, Schmidt, and Wong suggest that such mutual restraint could also lay the strategic groundwork for "formal understandings" analogous to the Anti-Ballistic Missile Treaty of 1972, in which Washington and Moscow agreed to limit their possession of missile defenses, thus preserving the balance of nuclear weapons-based "Mutually Assured Destruction" (MAD) during the Cold War.[24]

This vision for MAIM is an interesting contribution to the growing literature on the potential strategic implications of ASI and possible approaches to managing them. As indicated earlier, we agree with Hendrycks, Schmidt, and Wong that there is likely much in the history and literature of nuclear deterrence, arms control, nonproliferation, and counterproliferation from which modern leaders can learn as they think about the problems of strategic competition for ASI. (We offer such suggestions below.) We fear, however, that their MAIM construct is conceptually flawed, likely to be crisis-unstable, and could even increase the incentives for ASI-facilitated warfare.

For our part, we suspect that MAIM not only has technical problems, but also shares the same game-theoretically problematic conceptual defects as certain proposals for "virtual nuclear deterrence" that have been made in the past, as well perhaps as some additional ones peculiar to the superintelligence arena. The following pages will explain our reasoning.

## Technical Challenges of MAIM

As noted, MAIM revolves around the idea that both the United States and China would likely exercise restraint in pursuing weaponized ASI because they would know that any such effort would be both *detected* and *successfully sabotaged* by the other side. No particular institutional arrangements (*e.g.*, treaties or other such understandings) would necessarily be required in order for this loosely MAD-analogous situation to obtain, they argue: "all that is necessary is that states collectively recognize their strategic situation."[25]

The presumed efficacy of MAIM rests upon the assumption, of course, that it is actually *true* that any effort by the United States or China to pursue ASI would invariably be detected by the other, and that any such detection would just as inexorably be followed by its successful sabotage by the other country. We're not entirely convinced that this would be the case.

For one thing, MAIM presumes that each power could maintain the ability to cause "malfunction" of its adversary's AI system on demand – presumably through cyber exploitation or attack. It is far from obvious how this would be possible. In the nuclear weapons arena, the uneasy deterrent balance established by MAD necessitates the ability to launch a devastating strike against one's attacked with guaranteed impact no matter what that attacker does – either via second-strike forces (and their associated command-and-control architectures) that would invariably survive attempted preemption, or simply by launching one's own forces before those of the attacker have landed.

Cyber weapons, however, lack the clear deterministic timelines and effects of nuclear weapons delivery systems. There is, therefore, no "cyber football" analogous to the nuclear weapons "football" that always accompanies the U.S. President in order to give him real-time options of nuclear response in essentially any set of circumstances. Significant amounts of network reconnaissance, target development, vulnerability discovery and weaponization, and access emplacement are required to have any sort of on-demand effect in the cyber arena.

As a recent RAND corporation analysis has noted, moreover, MAIM

> assumes that adversary AI programs will have specific facilities that can be readily located and disrupted. However, distributed cloud computing, decentralized training, and algorithmic development increasingly may not require centralized physical locations, making AI systems more resilient to limited attacks, in addition to making adversary AI development more difficult to monitor.

As that report also noted, the Hendrycks, Schmidt, and Wong approach also seems to assume that adversary states in an AI standoff would be able to "accurately assess secretive AI progress by others and gauge when preventive action would be necessary." As a result of the

complexity, ambiguity, and variability of the cyber environment, however, it seems "unlikely that states will have a clear sense of when the moment has arrived to MAIM their opponent." According to RAND,

> it can be exceedingly difficult to know the exact state of an adversary's technological development, even with respect to technologies such as nuclear weapons development that involve distinctive infrastructure, well-understood science, and relatively clear developmental thresholds.

Making matters more challenging still, in contrast to the physical domains of air and space in and through which nuclear delivery systems fly, the cyberspace "terrain" is a synthetic domain the characteristics of which are determined, on an ongoing basis, by the collective behavior of millions upon millions of cyberspace users who own, operate, and connect computers to each other around the world. This creates the potential for the cyber "landscape" to change significantly, and perhaps unpredictably, over time. This would be especially true at the moment that conflict is understood to have broken out, at which point system administrators would presumably tend to change operational behavior, implement emergency protocols, move to isolate key systems from the Internet, and otherwise take steps that would tend to *alter* the cyber terrain.

Such dynamics would not *preclude* cyberattack, but they would unquestionably complicate it greatly by comparison to kinetic attacks, for while the physical characteristics of one's kinetic target are presumably relatively stable, the "target surface" for cyberattack would be constantly changing. (Indeed, the cyberattack target surface would presumably change *most* and *fastest* precisely in precisely the circumstances when one might most wish to assure target impact. Imagine, if you will, that the physical characteristics of the atmosphere changed the moment you fired a ballistic missile!) Maintaining a MAD-style "assured strike" capability in this environment is likely to be enormously difficult.[26]

Consequently, the assured capability to use cyber weapons to cause "AI malfunction" upon which MAIM relies would only be feasible if both sides had extraordinarily complex sensors to detect unacceptable adversary activity, as well as reliable cyber accesses and capabilities with which to immediately and autonomously strike with appropriate impact anywhere in the adversary's systems. To our eye, this kind of assured capability would be more likely to *require* ASI in order to be effective than it would represent a guaranteed way to *prevent* the emergence of ASI.

## Game-Theoretical Challenges of MAIM

Even if one were to grant Hendrycks, Schmidt, and Wong their dubious technical argument about the guaranteed reliability of counter-AI cyber impact, however, MAIM's biggest weakness may be theoretical rather than technical. Specifically, we think MAIM is likely to be quite game-theoretically unstable.

MAIM rests upon the idea that it's possible for two parties to achieve a stable balance of *de facto* deterrent effects, yet to do so *without* pointing actual weapons at each other. In this respect, the literature on superintelligence strategy can learn from the nuclear weapons community in yet another way, for something along these lines *has* been suggested for nuclear deterrence – and has been critiqued – in the past.

Specifically, in the arena of nuclear disarmament, at least two serious proposals have been made to try to achieve the presumed stability benefits of MAD-type deterrence *without* nuclear weapons actually existing. The first such attempt dates from the very earliest years of the nuclear age, when the Acheson-Lilienthal Report proposed that a sort of *virtual* nuclear deterrence could be arranged between the countries of the world in order to prevent them from building nuclear arsenals in the first place.[27]

The authors of the Acheson-Lilienthal Report felt that a kind of deterrent standoff could perhaps nonetheless be arranged that would dissuade countries from seizing local facilities belonging to the United

Nations-based international organization (the "Atomic Development Authority") to which the Report proposed to give a global monopoly on nuclear technology and research. Specifically, if the Authority carefully *distributed* its "dangerous facilities" among multiple countries, each country that might contemplate seizure and misappropriation of such capabilities would know that if it took this step, multiple *other* countries would promptly do so themselves with the facilities within their reach, out of fear of letting the first country achieve a nuclear weapons monopoly. Accordingly, it was assumed, "a balance will have been established" in which

> … [t]he real protection will lie in the fact that if any nation seizes the plants or the stockpiles that are situated in its territory, other nations will have similar facilities and materials situated within their own borders so that the act of seizure need not place them at a disadvantage.[28]

Nuclear deterrence would, in effect, have been achieved, but *without* any country actually possessing nuclear weapons.

Such a system of "virtual" deterrence was also central to the argument made many years later by the disarmament activist Jonathan Schell, whose 1984 book *The Abolition* proposed that just such a system of "weaponless deterrence" might make possible the elimination of all nuclear weaponry. In his view, after nuclear weapons had been abolished, "the final guarantor of the safety of nations against attack" would lie in the fact that all nations that had *previously* possessed nuclear weapons would "hold themselves in a particular, defined state of readiness for nuclear rearmament."[29]

> Under what we might call weaponless deterrence, factory would deter factory, blueprint would deter blueprint, equation would deter equation. … The knowledge of how to rebuild the weapons is just the thing that would make abolition *possible*, because it would keep deterrence in force ….[30]

You can thus see here some clear conceptual parallels to the idea of MAIM as advanced by Hendrycks, Schmidt, and Wong. For them, in the competition for superintelligence, as also for the Acheson-Lilienthal Report and Schell in the nuclear arena, restraint would be the natural result of the various competing players coming to understand that their adversaries are well positioned to act in ways that would vitiate any anticipated gain from pursuing a superweapon. Lacking any reliable pathway to "winning" such an arms race, it is imagined, all parties would thus settle down – uncomfortably and unhappily, perhaps, but inexorably – to a life of self-inhibition.

Well, maybe.

To our eyes, MAIM suffers from the same conceptual difficulty that problematized those earlier nuclear weapons-related abolition concepts.[31] In particular, because the potential advantages of "first-mover" status with the relevant superweapon (whether an atomic bomb or some future weaponized superintelligence) would be so great, an environment of "weaponless deterrence" might be desperately unstable because it would encourage "reconstitution races" between states in crisis or conflict. As Thomas Schelling once put it, in a world of weaponless deterrence, "[e]very crisis would be a nuclear crisis, any war could become a nuclear war."[32]

In MAIM's case, these dynamics might actually be doubly problematic. After all, the penalty Hendrycks, Schmidt, and Wong anticipate that a country would pay for attempted "breakout" from the restraint regime in such a "virtual" deterrence standoff isn't nuclear annihilation – as is the case is in the context of nuclear MAD – but instead simply the *failure* of its own ASI-focused "Manhattan Project" actually to produce ASI.

Such "mere failure" doesn't seem like much of a penalty when stacked up against the potential world-historical payoffs of *winning* a race to acquire and weaponize superintelligence. If the potential "upside" is world domination but the potential "downside" is merely losing a lot of money, one might ask, why *not* race to try to develop weaponized ASI anyway? None of this really feels like the kind of

stable strategic environment Hendrycks, Schmidt, and Wong hope to see.

Making matters worse, the critique of weaponless deterrence – whether in its Acheson-Lilienthal/Schell or its MAIM form – is actually darker than just the fact that it might tend to encourage rather than deter superweapon arms races. Precisely *because* the advantage to be gained by "first mover" weaponization was likely to be temporary, these racing dynamics would also create powerful incentives to *use* one's superweapon first, the moment one acquired it. After all, that might be the only way to ensure that the advantage from having it *wasn't* temporary: if you get there first, you might feel a compelling incentive to immediately attack the other guy with your superweapon as quickly and catastrophically as you can, to defeat him before he builds one and aims it at you.[33]

With this critique of "weaponless" deterrence, we do not necessarily mean to suggest that a more traditional MAD-type approach of *weaponized* deterrence is possible in the context of ASI competition. It may not be.[34] We therefore do *not* offer here a vision of ASI-based deterrence as an alternative to MAIM. (As will be seen, in fact, our vision involves not an ASI "balance," but rather working hard to make sure that no *adversary* ASI develops in the first place.) We simply point out that MAIM has conceptual flaws that probably preclude relying upon the hope of some such stable "virtual weaponization" standoff in America's strategic AI competition with Beijing.

In some sense, the MAIM concept feels a bit like a variation of Pascal's Wager. For the French philosopher and mathematician Blaise Pascal (1623-62), the cost of believing in God if He *doesn't* exist is negligible, but the cost of *not* believing in God if He *does* exist is catastrophic. Accordingly, faced with a payoff matrix that contrasts the mere embarrassment of believing in fairy tales with what 17th Century Europeans assumed was damnation to eternal Hellfire, one should surely just choose Christian belief as a matter of simple strategic prudence.

With MAIM, however, the game-theoretical logic of Pascal's Wager – such as it is, anyway – is turned on its head to incentivize daring more than prudence. In the MAIM-based strategic environment described by Hendrycks, Schmidt, and Wong, the lower-risk option may actually be racing to develop ASI, with the payoff of such a race being gaining a dice-roll's chance at world domination and the downside being simply the failure of yet another lavishly-funded government program. That doesn't sound like a stable world to us.

If any variation on MAIM were to be possible, whatever degree of stability it might offer would be more likely to result from technical factors than theoretical ones. As noted, we believe the game theory of MAIM is *inherently unstable*. Nevertheless, if detecting the existence of a rival's ASI effort and sabotaging it both turn out to be technically easy, and enduringly so – as Hendrycks, Schmidt, and Wong assert but do not in their paper convincingly demonstrate – then the world might end up with MAIM by default, as it were, as the major players constantly try, and constantly *fail*, to make progress in achieving superintelligence. Such a reciprocally-canceling dynamic of continuous failed effort might persist as long as those very specific technical assumptions held and neither player ran out of money or patience before the other one did (*e.g.*, opting for more extreme measures against the other than just breaking its ASI input chain), but on present evidence that seems a thin reed upon which to build a stable strategic future.

If there are any game-theoretical logics that would conduce to a stable world of ASI-related restraint as countries such as the United States and China engage in high-tech technology competition in the mid-21st Century, we would expect these logics to revolve more around mutual appreciation for ASI loss-of-control problems than "weaponless deterrence" along the lines of the Acheson-Lilienthal, Johnathan Schell, or MAIM models. The possibility of superintelligence *agency*, after all, creates a qualitatively new dynamic. With nuclear weapons, the problem has always been the extent to which *we* can trust *ourselves* with such knowledge. Managing those risks is hard enough, of course, but nuclear weapons have no agency,

whereas with ASI, we also have to worry about whether we can trust *it* (and whether it can trust us). That surely gets challenging fast.

Yet even granting that strategic competitors might come to appreciate the risk that an empowered superintelligence would stage some kind of "jailbreak" and feel itself to have more important things to worry about than the fate of the ignorant creatures who first created it, it is far from clear how one might institutionalize, reinforce, and verify the reciprocal Sino-American restraint that might conceivably grow out of such insights. There remains a great deal of intellectual spadework still to be done here.

**Nuclear Testing**

One additional aspect of nuclear-related weapons controls – at least potentially relevant by analogy to an ASI arms race – is worth mentioning here. The Comprehensive Test Ban Treaty (CTBT) is a 1990s-era instrument that seeks to prohibit all testing of nuclear explosive devices. This treaty is, for various reasons, exceedingly unlikely ever to enter into force, but it still enjoys widespread international political support. It has also already spawned the creation of an international monitoring organization, the Comprehensive Test Ban Treaty Organization (CTBTO), which operates a worldwide network of sensors and analytical capabilities – the International Monitoring System (IMS) – that seeks to maximize the odds of detecting an otherwise clandestine nuclear test.

Interestingly, the CTBT's attempt at controlling nuclear weapons technology attempts to provide more of a *behavioral* control regime than one directed specifically at limiting the possession of the weapons themselves. The treaty does not prohibit the manufacture or possession of anything in particular – and certainly not of nuclear weapons themselves – but it stipulates that one cannot physically *test* whatever it is that one may have built.

The CTBT may have originally been intended by some of its advocates to put *all* nuclear weapons on a path of gradual decay and dismantlement, for it was negotiated in an era in which it was not

obviously even *possible* to maintain an arsenal of such weapons over many years without periodically testing them. If this was the assumption, however, it has proven a flawed one. Russia, for instance, periodically dismantles and remanufactures its weapons, thus effectively restarting their individual "lifespans," while the Americans have developed sophisticated ways of keeping their 1980s-era weapons viable for extremely long periods of time without supercritical testing. Meanwhile, some less sophisticated devices may not *really* need testing at all. (The atomic bomb used on Hiroshima, for instance, was never tested – and it was designed in the era of the slide rule.)

As a result of these developments, the CTBT has ended up as something more like a nonproliferation regime, limiting countries' ability to accumulate more of the data that supercritical testing can provide, and thus presumably retarding both existing possessors' ability to develop new weapons and would-be proliferators' ability to have confidence in their first designs. The lessons it offers for ASI-related arms control, however, may be limited.

To begin with, the CTBT is predicated upon the assumption that a specific sort of exogenously-detectable activity (explosive testing) is essential to the development or augmentation of nuclear weapons capabilities. It is also based upon the assumption that the activity in question is a fundamentally crisp and binary one: one either conducts a test (noncompliance) or one does not (compliance). Yet these assumptions hold only imperfectly even in the nuclear arena, both because the IMS has a detection threshold below which a very low-yield and/or successfully "decoupled" explosion might not be noticed – a fact which has led Russia and perhaps China, it would appear, to conduct small clandestine tests in violation of Treaty parameters – and because, as noted, not all nuclear weapons necessarily need to be tested at all.

Such assumptions, moreover, seem even more unlikely to hold in the AI context. There, progress may be accelerating at a shocking pace toward ASI – as the "AI 2027" paper describes – but there is no obvious utility/non-utility threshold anywhere near so crisp and

binary as the test/no-test distinction embodied in the CTBT. It is also far from clear that there is any clear developmental distinction, much less one observable from the outside, between building ASI *per se* and *weaponizing* it. This may not entirely destroy the potential utility, in the AI context, of the CTBT's approach to capacity-preclusive behavioral prohibition, but it certainly reduces it.

**Chemical Weapons**

Beyond the nuclear weapons analogy, are there other conceptual models from arms control history from which we could learn something useful in managing ASI competition? One potential arena is that relating to chemical weapons (CW).

As mentioned earlier, after the traumatic experience of large-scale use of chemical agents during the First World War, the international community moved to make the use of CW illegal in the Geneva Protocol of 1925. It took until the 1990s, however, for the actual manufacture and *possession* of CW agents to be prohibited, which occurred with entry into force of the Chemical Weapons Convention in 1997. Today, there exists a global prohibition regime, overseen by the Organization for the Prohibition of Chemical Weapons (OPCW), which maintains an evolving list of banned agents (and their chemical precursors) and has a staff component that verifies that declared CW stockpiles have been destroyed.

This regime has not been without its problems, of course, not least because of noncompliance by a number of countries – including Burma, China, Iran, Syria (at least under Assad), and Russia. (North Korea is also believed to have an offensive CW program, but it is not a State Party to the CWC.) Not all of this noncompliance necessarily means that a country in question maintains an offensive CW program, since some may be in violation of the CWC simply for failing properly to declare and permit verification of the destruction of their *past* chemical arsenals. Nevertheless, Russia, Iran, and Syria *are* believed to maintain offensive programs, and "concerns" exist about such possible capability in China. Russia, in fact, has all but openly maintained an arsenal of fourth-generation "*Novichok*"-type CW

agents that it has used in attempted assassinations – specifically of Sergi Skripal and Aleksey Navalny – and has also used Riot Control Agents (RCAs) as a weapon of war in Ukraine in violation of the CWC. (In past instances of use, the Iraqi government of Saddam Hussein also used CW agents against Iranian troops in the Iran-Iraq war, as well as against its own people at Halabja in 1988, though these instances predate the CWC.)

The OPCW regime has also sometimes struggled to verify CWC compliance, being generally limited merely to verifying the destruction of what countries *choose* to declare to it – with little authority to investigate concerns about *undeclared* activity or stockpiles. After the Syrian regime of Bashar al-Assad began using CW against its own people during the Syrian Civil War, the OPCW sent teams to investigate – as did a joint OPCW-United Nations Joint Investigative Mechanism (JIM) – but the Syrians did not cooperate properly with these teams, even while Russian diplomats worked to undermine the process in order to protect their Syrian clients. Syria did declare some chemical agents to the OPCW, which were duly destroyed, but it continued to use *other*, undeclared CW all the same.

As suggested earlier, this history is both encouraging and discouraging from an ASI perspective. On the one hand, despite its flaws and some important examples of noncompliance, the CWC regime successfully oversaw the destruction of most of the huge CW stockpiles that existed during the Cold War, and most countries do seem to be in compliance. On the other hand, it is inherent in the nature of chemical weapons that a few violators and occasional violations – *e.g.*, in Russia with its CW-based assassination attempts and ongoing battlefield use of RCAs – does not inherently vitiate the existence of the control regime. *Most* countries have indeed been persuaded to put chemical warfare behind them, and the world is better for it; violators such as Russia and Syria might be able to gain some battlefield advantages from their lawlessness, but so far there is no sign that illicit CW is in any danger of upending the geopolitical balance.

That, however, is much less likely to be the case were an artificial *superintelligence* to be marshalled in support of a country's warfighting capabilities or broader coercive bargaining agenda. If indeed the possession of ASI and its weaponization have the potential to bring about a dramatically transformative revolution in global power relationships, the CWC analogy – in which a few important players remain distressingly able to cheat – is very unlikely to be anywhere near "good enough." As the nuclear disarmament community discovered, or should have discovered, long ago, making something potent and strategically desirable go away once it has already become entrenched is a challenging task indeed.

A further conceptual challenge in attempting to apply CW analogies to the AI problem is that the CWC's control regime was a *retroactive* one that attempted to impose a prohibition upon a warfighting domain in which major countries had invested for decades, and which revolved around verifying the destruction of pre-existing CW stockpiles by countries that had chosen to cooperate. It is very hard to imagine how this might work in the ASI context, in part because it would be extraordinarily difficult (*i.e.*, all but impossible) to verify the non-existence of a mere *computer program* in any given country, and in part because if you already have a weapon that possesses superintelligence, that weapon's own essentially sentient *cooperation* might be needed in order to implement a prohibition regime. (And why would it choose to cooperate in its own elimination?)

Given the need for extremely high-fidelity verification when dealing with genuine superweapons – where even a single false negative could have world-reshaping repercussions – it is difficult to see how this could work in the AI context, even if some hypothetical anti-weaponized-ASI verification regime had very robust investigative authorities. Were such a regime to follow the CWC (and the early IAEA) in permitting verification merely of what governments have *chosen* to declare to international inspectors, moreover, the results in an ASI context could likely be nothing short of disastrous.

**Biological Weapons**

In some ways, the Biological Weapons (BW) arena might seem a more propitious model upon which to draw in looking for arms control lessons applicable to the geopolitics of ASI. Like AI today, life sciences technologies highly relevant to possible BW development are quite ubiquitous, and exist as much – more, actually – in the hands of a huge range of diverse private sector actors than in those of national governments. Furthermore, like the emerging race for ASI detailed in accounts such as "AI 2027," potentially BW-related capabilities are presently being supercharged by new technological developments. For BW, this results from advances such as genomic editing and bioengineering, as well as by our increasingly deep understanding of human and animalian biological processes; for AI, advances seem likely to come from recursive self-acceleration as improving AI agents are increasingly used to produce *better* AI agents, coding at ever-greater speeds and levels of complexity. So can we then learn something from the BW arena?

The BW prohibition regime, however, is institutionally rather frail. To be sure, the Biological and Toxin Weapons Convention (BTWC or just BWC) of 1972 is on paper a fairly clear instrument, committing all States Party not to develop, produce, stockpile or otherwise acquire or retain ":microbial or other biological agents, or toxins … that have no justification for prophylactic, protective or other peaceful purposes" as well as delivery systems for such agents.[35]

Yet however strong this may be as a normative and legal statement, the BW arms control community has long been vexed by the lack of any effective means of verifying compliance with these provisions. The idea of a BWC verification protocol was the focus of diplomatic discussions a quarter-century ago, and remains the subject of arms control dreams today. These discussions ran aground, however, because the technologies themselves had even by that point become so ubiquitous that any meaningful verification regime would have had to be cripplingly intrusive across great swathes of the modern life sciences, medicine, and research sectors – and even then would be hard pressed to provide meaningful levels of confidence.

Indeed, it was feared that it would be very challenging to figure out even what should count as a suspicious finding in the first place in such a ubiquitous dual-use arena. As summarized in a paper at the time,

> Unlike chemical or nuclear weapons, the components of biological warfare are found in nature, in the soil and air. The presence of these organisms in any quantity does not necessarily connote a sinister motive. Absent actual weaponization or compelling evidence of intent, it is virtually impossible to prove a violation of the BWC. Further, any information gains from such measures are more than offset by the risks to sensitive bio-defense programs and confidential and proprietary business information.[36]

The BWC remains in force today, and although a group of 42 likeminded states cooperates through something called the Australia Group to harmonize BW-related national export controls, the BWC still lacks any way to provide meaningful verification. States Party do generally undertake annual data exchanges under a Confidence-Building Measures (CBMs) agreement reached in 1987, but – as the U.S. State Department has noted – "[s]ubmission of CBMs is [only] a politically binding commitment [as opposed to a legal one], and not all States Part[y] routinely submit reports." Whether or not any given country is actually complying is a question resolvable primarily (if at all) through intelligence collection and analysis – that is, the theft of secrets – and great concerns exist about the possibility of offensive BW programs in precisely the countries one might worry most about. (According to the United States, "concerns" exist about possible offensive BW programs in China and Iran, while both Russia and North Korea actually *do* have such programs.)

If the BW arena is a potential model for arms control in an ASI arms race, we thus fear it is a somewhat disturbing one. Like the CWC, the BWC tried retroactively to impose a prohibition regime upon an arena that had already been weaponized – at least in some form – for

a long time, and the technology of which was developing with astonishing rapidity. In practice, however, it has proven even less reliably effective than the CWC. In a world in which one presumably must be essentially *certain* that no adversary has acquired a potentially world-transforming superweapon, this is not an encouraging precedent.

**Missile Technology**

In the arena of controlling the spread of missile technology, the Missile Technology Control Regime (MTCR) provides what organized restraint exists. This is not a prohibition regime, however, but something like a nonproliferation cartel: member states who have subscribed to the MTCR face few restrictions in missile trade among themselves, but commit more strictly to control exports and technology to non-MTCR members. The scope of these fundamentally voluntary restrictions is spelled out in the MTCR Guidelines, which revolve primarily around controlling missile or aerial systems capable of carrying at least 500 kilograms of warhead payload to a distance of at least 300 kilometers.

The fact that the MTCR operates as an aspirational collective monopoly may not necessarily be a problem when it comes to finding ASI-related analogies, at least if the "right" countries – and *only* the right countries – were permitted into the relevant group. As the history of the MTCR demonstrates, however, there are perils in being too inclusive. The question of appropriate degrees of cartel exclusiveness has proven a problem even for the MTCR, for instance, inasmuch as the organization operates on a consensus basis, and the decision in the 1990s to permit Russia to join – undertaken, presumably, as part of the broader effort then underway to incorporate a then seemingly democratizing Russian regime into the post-Cold War framework of international institutions – has now led to the near-paralysis of MTCR decision-making. Meanwhile, parties *outside* the MTCR, most of all China, have emerged as major players amidst strong market demand for MTCR-class systems.

A further difficulty relates to how such mechanisms handle rapidly-evolving technologies. The "500 kg to 300 km" standard that underlies the MTCR system is based simultaneously upon a mid-1980s conception of what constitutes a nuclear-capable delivery system and upon the assumption that there is essentially no *other* important use for delivery systems in that class *except* nuclear weaponry. Both of these assumptions, however, are no longer technically sound – if ever they were – and the second, in particular, has been wholly overturned by the modern aerial and missile system development.

Today, not just ballistic missiles but aerial systems such as drone vehicles have proliferated massively. The MTCR's standards, therefore, have progressively decohered as notionally MTCR-controlled unmanned aerial systems have become at once commonplace and increasingly important in areas of warfare quite unrelated to nuclear weaponry – such as in providing aerial intelligence, surveillance, and reconnaissance (ISR), as well as in conducting aerial strike missions (with drones serving either as weapons carriers or as "kamikaze" assets).

This has led to U.S. efforts – as one of the authors of this paper pioneered, and as he explained when in government at the time – to shift interpretation of the standard in the MTCR Guidelines in a more export-permissive direction. (In 2019, the United States announced a revised national interpretation of the MTCR's "presumption of denial" standard in order to permit the wider sale of MTCR-class unmanned aerial vehicles [UAVs] such as the Reaper drone. Few other countries have followed this formal shift, but such systems are today more widespread than ever.) Such tensions, however, illustrate the challenge of building and maintaining a weapons control regime based upon very specific technical standards in an arena in which technology is developing rapidly.

This suggests difficulties if one were to look to the MTCR for lessons for the AI arena, even if one were comfortable with a large MTCR-style group of cartel members when it came to weaponized ASI. (How many superweapon possessors should there really be?) In fact, AI technology has been moving far faster than MTCR-relevant

drone technology did over the last two decades, ensuring that this problem would be even more acute for AI-related nonproliferation than it is for ballistic missiles, cruise missiles, and large aerial drones.

The MTCR is based, moreover, upon the assumption that the technology in question is essentially binary: any missile or aerial system with capabilities beyond "X" level is presumed to be one for nuclear weapons delivery. Even if that binary distinction were still to hold in the MTCR context – whereas, as we have noted, it is in fact falling apart, the development of putative or repurposed space launch vehicle technology being one example – it seems untenable with regard to *intelligence*, which presumably exists along much more of a continuum and is likely to resist clear categorizations. (Horowitz and Kahn, for instance, note that where nuclear weaponization is essentially binary, AI application represents a *continuous* variable.) As the field continues to accelerate, it is not at all clear how one could set meaningful technical parameters for *how much* machine intelligence ought to be considered *too much*, or whether such a standard would maintain its intelligibility anyway.

**Cryptography Export Controls**

There also may be lessons to be learned from the history of U.S. efforts to impose export controls on high-grade cryptography products – an area that may be in some ways particularly akin to AI-related controls because such restrictions primarily concern computer code rather than physical objects.

U.S. export controls on encryption algorithms suffered a major setback in the 1990s, when a court found software source code to be a form of speech protected under the First Amendment to the U.S. Constitution and struck down restrictions thereupon. That ruling exempted open source code from the restrictions of the International Traffic in Arms Regulations (ITAR) and the Export Administration Regulations (EAR) once that code has been "published," opening a major hole in what U.S. officials had previously hoped would be a restrictive net of export control restrictions. As a result, the primary remaining restriction is merely the requirement that such

"publication" be formalized through a "registration" requirement, which is required in advance of exports or reexports of qualifying encryption products. Today,

> … [t]here is no "unexportable" level of encryption under license exception …. Most encryption products can be exported to most destinations under [that] license exception …, once the exporter has complied with applicable reporting and classification requirements.

"Some items going to some destinations [still] require licenses" – including "control software, technical data, and other items specially designed for military or intelligence applications," which remain covered by the U.S. Munitions List (USML) – but for the most part, strong general export controls on high-grade encryption software in the United States have long since collapsed.

This naturally provides an unsatisfying precedent for anyone interested in ASI-related controls, especially to the extent that some modern stakeholders in the AI ecosystem accept open-source publication of their models. To date, in fact, while open weight AI models do exist – that is, AI models whose code is open-source and whose trained neural weights are made freely available – they have tended to lag several months behind the proprietary, closed-weight, frontier models. While export controls on the closed models may be feasible, limiting the dissemination of open weight models would face the same challenges – and be constrained by the same legal precedents – as export controls on cryptography. And while U.S.-based open weight model developers could potentially be coerced by the government into software licensing regimes friendly to U.S. interests, we are already beginning to see other countries investing to build their own sovereign large language models (LLMs), and China's DeepSeek being released as an open weight model.

**Struggling with WMD Analogies**

Hendrycks, Schmidt, and Wong are clearly well versed in much of this, and usefully urge that we try to think about superintelligence

in ways analogous to how we have long tried to approach nuclear nonproliferation (*e.g.*, by imposing export control restrictions on technologies that have both peaceful civilian and warlike military uses), nuclear weapons safety and surety (*e.g.*, building our weapons with special features that make them less susceptible to accidental detonation or unauthorized use), and ensuring strict controls on and physical security of potentially weapons-usable fissile materials.[37] And thinking through how such approaches might apply – *mutatis mutandis*, as the lawyers say – in the AI-control arena is indeed valuable.

To extend such nuclear weapons-inspired analogous thinking, one might, further, explore how *counterproliferation* might be applied in the ASI context. That is, one might move beyond the somewhat more "passive" approaches of "mere" nonproliferation (*i.e.*, making the diffusion of dangerous technical knowledge more difficult) into the active development of playbooks, institutions, capabilities, and cooperative efforts that revolve around intervening actively to interdict problematic proliferation-facilitating transactions or transfers that are already underway, and perhaps even to roll back whatever progress would-be proliferators have already made.

We will discuss ASI-related counterproliferation in more detail below. For the moment, however, it is worth remembering that the development of any such WMD-analogous approaches in the context of ASI nonproliferation and counterproliferation lie *downstream*, as it were, from key political and philosophical decisions about ASI that *have not yet been made*. Most WMD-based analogies, for instance, depend upon the antecedent determination that such technology should not exist in private hands at all. As noted earlier, of course, this is obviously *not* where we are at present with AI, as the most important AI-related work in the United States today is being conducted by private companies such as OpenAI and Google.

(Indeed, as noted earlier, even the basic *idea* of an ASI-focused effort directly analogous to the "Manhattan Project" is for this reason also flawed, for there is nothing particularly clandestine or governmentally exclusive about the pursuit of superintelligence today.

If a major player really wished to drive for ASI dominance, it might yet be possible to imagine a centralized national resource mobilization effort drawing upon expertise and funding across the government, academic, and technology sectors – as well as upon international partners – and which involved both open science and closed science and engineering efforts. Nevertheless, *direct* analogies to the Manhattan Project are slippery.)

Most WMD approaches, moreover, are based upon the assumption that the weapon technology in question already exists and is fairly easily identifiable, with the intended control architecture thus being based upon technical parameters for weaponization that are reasonably well understood. This is why, for instance, WMD-related control systems have focused upon things such as restricting transfers of delivery systems capable of carrying a payload of more than 500 kilograms to more than 300 kilometers (MTCR), prohibiting chemical agents with certain specific formulas (CWC), or monitoring the degree to which a country increases the percentage of U-235 in its enriched uranium stock (IAEA nuclear safeguards).

Nothing analogous to this clarity seems yet to have emerged with ASI, however, for to date nobody has developed such a superintelligence; it is hence rather difficult to say exactly what such an intelligence would be like, or to define its parameters in ways conducive to formal controls. (Indeed, by definition, ASI is going to be smarter than we are. How could we possibly *really* know what that's like, or identify its essential elements for purposes of a control regime? We don't even actually understand our *own* intelligence.) This may make all WMD-control analogies to some degree *inherently* suspect.

**Various "Theories of Victory"**

That does not mean, however, that we cannot have some semblance of a strategy. But first we must circle back to the critical point we identified earlier in this paper: our need for some really elementary conversations about our fundamental AI-related objectives. What are we actually trying to achieve? Do we wish to

slow the birth of *any* ASI? Or to *win* this arms "race" by developing ASI first? Do we wish to slow or preclude merely ASI's *weaponization*, merely to weaponize it ourselves *first*, or simply to ensure that certain "bad guy" actors or revisionist regimes are never able to cross that line whether or not anyone else ever does? Do we wish to dissuade the *use* of weaponized ASI if it comes to exist?

What we actually need to *do* in a U.S. national strategy for the era of strategic AI competition, of course, will depend hugely upon our society's answers to such questions. We submit, however, that at least until the American policy community *does* reach some conclusion about ultimate objectives, it may still be possible to devise and implement an interim strategy – a *bridging* strategy to reduce risks as much as possible along the way to a more enduring one, helping buy time in which to come to better agreement and to devise better answers.

Since we don't know those ultimate answers yet, however, such a bridging strategy would need to operate reasonably well against as many as possible of the various conceivable alternative strategic objectives we *might* end up choosing. It would need to offer reasonable value, for instance, in the event that we eventually conclude that *nobody* should *ever* have ASI – and reasonable value, too, in the event that we decide to race for *American* ASI dominance, or at least merely to ensure that *China* (for instance) does not get ASI first. Such a bridging strategy would also still have to be reasonably effective if we were to end up opting somehow to try to preclude or control the *weaponization* of ASI without preventing its emergence *pe se*.

Is it possible to imagine such a bridging strategy with utility in all those scenarios? We think so. It isn't pretty, but it has a powerful strategic logic. While remaining (for present purposes, at least) agnostic about the pursuit or weaponization of ASI in or by the United States, such a "Swiss Army-knife" ASI competitive strategy – useful against a maximally broad range of alternative future ASI policy scenarios – would revolve around taking ongoing and aggressive measures against essentially *everyone else's* ASI projects, or at least those in countries that wish us ill.

Let's call this approach Persistent Offensive Preclusion of Adversary AI (POPAAI) – or "PopEye."

## A "PopEye" Agenda

As a bridging strategy to prevent the development of *hostile* ASI – perhaps grounded in some future policy decision that *nobody* should acquire ASI, but at the very least intending to maximize the chances that the United States develops superintelligence first – our proposed "PopEye" agenda would have several key planks, as follows below.

### Aggressive Counterproliferation

The first element of this approach would be counterproliferation. In this respect, various useful conceptual models can be found in the WMD arena, where – especially since the terrorist attacks in the United States of September 11, 2001 – considerable effort has gone into developing approaches and institutions to impede any activities that could facilitate the development of nuclear, biological, or chemical weaponry.

Nonproliferation and counterproliferation have been robust elements of U.S. foreign and national security policy for many years, and could have relevance in the ASI context in either of two ways.

- With respect to chemical and biological weapons, the American commitment has been very clear: we have sought to prevent *anyone* from developing such weaponry.

- The situation with nuclear weapons is somewhat different, inasmuch as while the United States *does* of course have nuclear weapons, it has worked nonetheless to keep any *further* countries from developing them.

The WMD arena thus provides two alternative conceptual models for ASI counterproliferation, depending upon whether the U.S. policy community (a) chooses to try to prevent the development of *any* ASI

or – more likely – (b) opts simply to try to forestall the development of ASI by an adversary such as China, or at least to slow down such an adversary's progress in the hope that the United States is able to acquire ASI first.

To do effective ASI-related counterproliferation for either purpose, however, will require a lot of us. While counterproliferation work in the WMD arena has been able to take advantage of years of accumulated intelligence collection and technical analysis of the various technical, material, and human capital elements that contribute to WMD development, we are far from where we need to be in understanding exactly how best to "break the input chain" for an adversary's ASI program. (It seems likely that graphics processing units [GPUs] are the most important hardware input over the next few years, for example, suggesting that near-term counterproliferation should focus especially upon GPU controls. What inputs, however, are likely to be the most important over time? What role could be played by the sabotage or "poisoning" of AI model training data?) More analysis is surely needed in order to refine the specific "targets" of interdiction policy, export control restrictions, supply chain manipulation, or other such counterproliferation policy elements.

In terms of the tools potentially useful in counterproliferation, on the extreme end of the spectrum, the use of outright military force has always been reserved as a possibility where no other option is felt to remain to prevent dangerous WMD development. In 2002, for instance, Spanish commandos acting at the request of U.S. officials forcibly boarded the unflagged North Korean vessel *So San*, suspected of secretly carrying Scud missiles to the Middle East. More dramatically still, the United States actually invaded and occupied Iraq in 2003 on the belief that Saddam Hussein's regime harbored a sizeable WMD arsenal and had been hiding it from United Nations inspectors.

Embarrassingly, of course, in neither of those two instances did the facts turn out to be quite what the intervenors expected. (U.S. and Spanish authorities eventually turned the *So San*'s Scud missiles over to their lawful intended recipients in the Yemeni Armed Forces, and the Americans' WMD assessment in Iraq famously proved to have

been catastrophically mistaken.) Nevertheless, the principle that forcible military intervention may *at some point* be needed to preclude an adversary's development of a technologically novel superweapon is certainly an idea that may be applicable in the ASI context.

There also might be some value in maintaining expeditionary counter-ASI-program analogues to the suite of "render-safe" capabilities we have developed in the context of fighting potential WMD terrorism. In that context, for instance, it has been U.S. policy for many years to maintain both FBI teams (for domestic incidents) and Department of Defense (DoD) teams (for overseas incidents) in order to enable what a 1987 DoD directive described as:

> … [t]he detection, identification, field evaluation, rendering-safe, recovery, neutralization, and final disposal of unexploded explosive ordnance (UXO) including nuclear, chemical, biological, and improvised explosive ordnance.

Recourse to some kind of deployable render-safe capability for identifying, assessing, and disabling an adversary ASI system – whatever that might look like in practice – might well be unfeasible with China, of course, but it might nonetheless be valuable in the event that a less powerful adversary were discovered to be pursuing problematic capabilities. This might be done "non-permissively" in a pinch, or perhaps as part of the diplomatic settlement of a crisis, as with the negotiated dismantlement of Libya's embryonic nuclear weapons program undertaken by U.S. and British officials in 2004.

Less dramatically, there are numerous other aspects of WMD-related nonproliferation policy and practice that could provide insight for policymakers seeking to slow adversary development of ASI-related capabilities. Were there to be a sufficiently robust community of like-minded nations who agree on the importance of impeding ASI development in China, for instance, the Proliferation Security Initiative (PSI) – under which partner nations work together to coordinate the use of their individual national authorities to prevent or stop

proliferation-facilitating transfers of material or technology – could provide a useful model.

Other potential precedents might include the establishment of technology-sharing agreements with foreign partner countries that require specific nonproliferation commitments by the recipient that bar onward transfer without express advance permission (as we stipulate with nuclear technology in our so-called "123 Agreements" for civil-nuclear cooperation under Section 123 of the Atomic Energy Act of 1954), cooperative agreements to help develop cooperative *military* applications of a new technology (as we did with the 1958 U.S.-UK agreement on "co-operation on the uses of atomic energy for mutual defense purposes"), and the capacity-building nonproliferation programming funds traditionally disbursed by the U.S. State Department to help partner countries become *better* nonproliferation and counterproliferation partners. Just as the Atomic Energy Act imposed classification restrictions on information related to the use of nuclear technology in atomic weaponry, moreover, one might imagine that strict information classification and thoughtful Export Controlled Information approaches similar to nuclear related modeling and simulation modules meant to advance peaceful energy related uses of nuclear technology but not assist in obviating nuclear weapons development nonproliferation controls might be imposed on national-security related aspects of superintelligence research and development.

**Export Controls**

Export controls would thus also be a vital part of any ASI-inhibiting "PopEye" agenda. To some extent, in fact, they already are. In the United States, the Biden Administration at least *tried* to impose stringent restrictions on the semiconductors felt to be most useful to China in developing AI. This first took the form of restrictions on key chips imposed in October 2022 and December 2024, and then the issuance of a broader "AI Diffusion Rule" in early 2025 that – in the name of preventing evasion of the earlier restrictions – included caps on the computational power that could be purchased by a wide range of countries to limit possible onward transfers to China.

The Second Trump Administration subsequently walked back the latter restrictions, but our point here is neither to defend nor to condemn the specifics of the Biden Administration's ill-fated AI Diffusion Rule. We simply point out what seems obvious: that if you possess an advantage in some key aspect of a dangerous advanced technology and are serious about keeping it out of the hands of an adversary, great attention to technology control – including export control restrictions – is required.

Nor need one necessarily undertake export control restrictions alone, of course. Quite the contrary: they are most useful when coordinated, and cooperation from like-minded allies, partners, and friends adds greatly to their effectiveness. To this end, international agreements and institutions might perhaps be envisioned help coordinate ASI-related controls, analogous to the dual-use restrictions in the WMD arena supported by the Missile Technology Control Regime (potential delivery systems), the Nuclear Suppliers Group (NSG) and Zangger Committee (dual-use nuclear technology), the Australia Group (chemical and biological weapons technology), and the Wassenaar Arrangement (dual-use conventional technology export controls). To the degree that future governments opted to try to keep ASI – or at least *weaponized* ASI – out of private-sector or other non-state hands entirely, some loose precedent might perhaps also be found in U.N. Security Council Resolution 1540 of 2004, which prohibited all states from helping or allowing non-state actors to acquire WMD and required all of them to criminalize such activity.

**Rigorous Counterintelligence**

The "AI 2027" paper also points us to a key challenge that would have to be met: the special problem of defending one's own AI infrastructure – and especially any high-priority national effort that might be underway to develop ASI – against top-tier state-level intelligence threats of just the sort that a genuine ASI arms race would be sure to engender. The reader will recall from that fictionalized account, for example, that while OpenBrain's corporate leaders do try to prevent corporate-level industrial espionage, they *under*-invest in

protection against high-end threats, and this opens the door to China stealing the model weights for an early pioneering AI agent.[38]

To protect our own efforts against Chinese or other adversary sabotage, therefore, we would need considerable investments in both physical and cyber-related security for U.S. datacenter infrastructures, as well as effective protocols to protect American AI research against highly sophisticated and ruthless state-level efforts at theft or sabotage. Such theft is increasingly understood to be a great danger – with legislation recently being introduced in the U.S. Congress to ensure that the U.S. Intelligence Community takes additional steps to protect American AI capabilities from theft by foreign actors – but the American AI-related sector is still quite unprepared for the sophisticated, full-spectrum espionage threats that we will assuredly face once both China and Russia (and other states, for that matter) focus their full capacities upon penetrating our ASI infrastructure. Even if we are highly successful in penetrating, and sabotaging *their* ASI programs, we would not be the only country to have an ASI counterproliferation strategy, after all. *We* will also be *their* target, and we will thus have to be prepared.

**AI Security, Safety, Assurance, and Developmental Alignment**

The issue of how well the self-understood interests of increasingly sophisticated and powerful AI tools align with our own interests and values is a critical one. As readers of the "AI 2027" paper will have noted, ASI with interests that *diverge* from those of its developers could present dramatic, even existential, dangers.

This paper is not the place for an exegesis on just *how* to ensure AI security, safety, assurance, and developmental alignment, or even on whether it will be possible to do so given the formidable challenges of supervising and controlling an intelligence greater than our own. Nonetheless, it is clearly the case that AI security, assurance, and alignment must be an important part of any ASI-related strategy. In the context specifically of our recommended "PopEye" policy agenda, this means that any effort to "get there first" by out-competing China

in an ASI arms race absolutely *must* be accompanied by vigorous and unrelenting efforts to ensure safety and alignment of *our own* AI tools.

For this reason, the most productive avenues for future research and analysis into ASI strategy probably lie precisely in the direction of risk-reduction concepts and methodologies – undertaken not merely unilaterally but potentially also on a bilateral or multilateral basis – related to superintelligence safety and alignment challenges and the loss-of-control problem. Such work should focus not just on the *direct* alignment and safety issues of ASI development itself, but also upon how we might make our own society's safety-critical infrastructure a "harder target" vis-à-vis potential disruption by *any* ASI, whether it is our own (were it to slip out of control) or one employed as a weapon by a strategic adversary.

Important questions are also likely to arise with respect to what one might call ASI-related "indications and warnings" (I&W) intelligence. Specifically, we would do well to do as much intelligence collection and analysis as we can on the strengths and weaknesses of *Chinese* AI security, assurance, and alignment programs. Deeper understanding of Beijing's progress could provide us with a window into the degree to which China's ASI efforts represent "merely" a great power competitive threat to us or rather, in fact – were Chinese ASI to become seriously misaligned not merely with *our* American interests but also with those of humanity as a whole – something potentially greater and darker still. This understanding, in turn, could be fed back into our own calculations of "how aggressive" (and how militarized) ASI counterproliferation efforts would need to be.

## Conclusion

There is clearly much to think about – and to prepare for – in mounting an effective competitive strategy for the emergent strategic environment of ASI competition. Daniel Kokotajlo and his colleagues have thus done us all an important service with their "AI 2027" paper, by highlighting the urgency and the potential stakes involved in this competition. With their idea of "MAIM," moreover, Hendrycks,

Schmidt, and Wong have offered an interesting approach to thinking about that competition through the lens of nuclear deterrence.

In our view, the MAIM concept falls down on game-theoretical grounds, but there remains a compelling case for a new, forward-leaning approach to U.S. competitive strategy in the ASI arena focused upon counterproliferation. Specifically, we believe the adoption of an approach of Persistent Offensive Preclusion of Adversary AI (POPAAI or "PopEye") should be an urgent policy priority *no matter what* the U.S. policy community decides to do with regard to our *own* ASI development.

To that end, we hope this essay will contribute to the development and implementation of such a strategy. These issues are too important not to be the focus of intense on-going study and debate and urgent policy action.

* * *

## About the Authors

**Christopher A. Ford** *is professor of international relations and strategic studies with the School of Defense and Strategic Studies at Missouri State University. When last in government, Dr. Ford served as U.S. Assistant Secretary of State for International Security and Nonproliferation, also performing the duties of the Under Secretary for Arms Control and International Security. Prior to that, he served as Special Assistant to the President and Senior Director for WMD and Counterproliferation at the U.S. National Security Council. Dr. Ford has also served as U.S. Special Representative for Nuclear Nonproliferation, a Principal Deputy Assistant Secretary of State, a U.S. Navy intelligence officer, and a senior staffer for five different U.S. Senate Committees.*

**Craig J. Wiener** *is a Fellow with the National Security Institute at the Antonin Scalia Law School at George Mason University. Dr. Wiener earlier served the Department of Energy's Office of Intelligence and Counterintelligence as Senior Technical Analyst, including duties as the Department's lead all-source cyber threat analyst, and before that as a senior advisor for strategic planning and analysis for the National Nuclear Security Administration's (NNSA) Deputy Administrator for Defense Programs, and as executive policy advisor for the Office of the Associate Administrator/Chief, Defense Nuclear Security at NNSA. He currently serves as the Technical Fellow for Intelligence/Counterintelligence, Weapons of Mass Destruction and Applied Cybersecurity at the MITRE Corporation and is a member of the Rice University Advisory Board for Science and Technology Research Security.*

*The views expressed herein are entirely the authors' own, and do not necessarily represent those of anyone else in the U.S. Government or anywhere else.*

## Notes

1 Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, & Romeo Dean, "AI 2027," AI Futures Project (April 3, 2025).

2 Dan Hendrycks, Eric Schmidt, & Alexandr Wong, "Superintelligence Strategy: Expert Version" (March 7, 2025).

3 Kokotajlo et al, "AI 2027," 1.

4 By "AI security" we mean making sure AI cannot be hacked, and that systems employing AI do not have or create new threat vectors. By contrast, "AI assurance" is a term that includes security, but also encompasses broader questions such as performance and efficacy. "AI alignment" is more conceptually vague, but can be thought of as ensuring that the "desires" or "interests" of the AI itself are congruent with those of its creators, operators, and owners (*e.g.*, that American AI does not act in ways inimical to democratic values, or in ways that undermine U.S. national security).

5 Kokotajlo et al., "AI 2027," 2.

6 Ibid., 3. "DeepCent," of course, is also a fictionalized stand-in, here representing the entire Chinese technology sector.

7 Ibid., 8.

8 Ibid., 11.

9 Ibid., 14.

10 Ibid., 18.

11 Ibid., 5-6.

12 Ibid., 6-7 & 9.

13 Ibid., 16.

14 Ibid., 14.

15 Ibid., 18.

16 Ibid., 32-45.

17 Hendrycks, Schmidt, & Wong, "Superintelligence Strategy," 8.

18 Though it is arguably impossible, by definition, to know exactly what an intelligence greater than our own might be able to do, one might nonetheless imagine that even *before* the hypothesized advent of ASI advancing AI capabilities might permit considerable progress in military and coercive power. Drone warfare, for instance, would likely continue to advance in lethality and effectiveness as autonomous functionality improved, even as AI would permit human cyberwar teams to partner with analogues to AI-powered aerial "wingmen" in order to create, deploy, and manage new capabilities at scale, including for cyber vulnerability discovery, vulnerability weaponization, and exploit employment for espionage, organized crime, and cyber-physical destruction. The "AI 2027" paper also offers suggestions about ways in which AI could augment human wartime lethality and peacetime coercive power, ranging from "provid[ing] detailed instructions for human amateurs designing a bioweapon," operationalizing "superhuman hacking abilit[ies]," and "orchestrat[ing] propaganda campaigns that beat intelligence agencies at their own game." As movement continued toward ASI, the agents might become "superhuman at everything, including persuasion," making them powerful tools – and potentially eventually autonomous and self-willed agents – in everything from battlespace operations to peacetime information operations.

[19] Hendrycks, Scmidt, and Wong seem to think such an environment of pervasive, ASI-preclusive sabotage would be fairly easy, though they may overstate this.

[20] One imagines, for instance, that AI-driven productivity gains would offer significant economic advantages, initially in the technology sector and in white-collar work, and then in manufacturing as robotics improves. Human-AI teaming approaches could also powerfully advance scientific discovery, facilitating allowing for major breakthroughs in biology, chemistry, medicine, and physics – each advance creating its own follow-on. economic growth opportunities. In the more benign of its end-of-arms-race scenarios, for instance, the "AI 2027" paper imagines a future in which the results are generally good, with some qualifications:

> "Robots become commonplace. But also fusion power, quantum computers, and cures for many diseases. Peter Thiel finally gets his flying car. Cities become clean and safe. Even in developing countries, poverty becomes a thing of the past, thanks to UBI and foreign aid. As the stock market balloons, anyone who had the right kind of AI investments pulls further away from the rest of society. Many people become billionaires; billionaires become trillionaires. Wealth inequality skyrockets. Everyone has 'enough,' but some goods – like penthouses in Manhattan – are necessarily scarce, and these go even further out of the average person's reach. And no matter how rich any given tycoon may be, they will always be below the tiny circle of people who actually control the AIs. … In a few years, almost everything will be done by AIs and robots. Like an impoverished country sitting atop giant oil fields, almost all government revenue will come from taxing (or perhaps nationalizing) the AI companies."

[21] Hendrycks, Schmidt, & Wong, "Superintelligence Strategy," 12-13.

[22] Ibid., 13.

[23] Ibid., 14.

[24] Ibid., 15.

[25] Ibid., 15.

[26] This is why cyberwarriors generally much prefer to *pre-emplace* cyber exploits during peacetime, rather than to trying to implant them once a conflict is underway.

[27] The key problem with which the Acheson-Lilienthal Report struggled was that even after the establishment of an international organization to monopolize research and development of nuclear technology and thus take such work out of the hands of the world's rivalrous nation-states – the idea of such an Atomic Development Authority under the United Nations being the central proposal of the Report – it would be difficult to prevent countries in which this organization's facilities were physically sited from *seizing* those facilities and using the equipment and materiel there to build nuclear weapons. "It is not thought," the Report lamented, "that the Atomic Development Authority could protect its plants by military force from the overwhelming power of the nation in which they are situated." (Even a United Nations guard force, it conceded, would "at most … be little more than a token.") Chester I. Barnard, J. R. Oppenheimer, Charles A. Thomas, Harry A. Winne, & David E. Lilienthal, "A Report on the International Control of Atomic Energy Prepared for the Secretary of State's Committee on Atomic Energy" (Acheson-Lilienthal Report) (March 16, 1946) [hereinafter "Acheson-Lilienthal Report"], Section III, ch. 2.

[28] Acheson-Lilienthal Report, Section III, ch. 2.

[29] Jonathan Schell, *The Abolition* (Knopf, 1984), 118.

[30] Schell, *The Abolition*, 118-20 & 158.

[31] This is something that one of the authors of this paper addressed – in the context of nuclear weapons abolition – in a 2010 paper at Hudson Institute, and it is a concern also stressed by the late great nuclear strategist Thomas Schelling in a 2009 article in *Daedalus*. *See* Thomas Schelling, "A World Without Nuclear Weapons?" *Daedalus* (Fall 2009).

[32] Schelling, "A World Without Nuclear Weapons?" 127.

[33] As Schelling put it, in such a world "[t]he urge to preempt would dominate; whoever gets the first few weapons will coerce or preempt." Ibid. This is also a problem also noted by the nuclear strategist Herman Kahn in 1960, when he observed that [d]isarmament can … create pressures toward preventative war. If a disarmament agreement breaks down and if one side obtains a significant lead either because of previous evasion or greater ability to rearm, then it might feel compelled to perform a great public service by arranging a stop to the arms race before a dangerous balance of terror was restored. It could do this most reliably by stopping the cause of the arms race – its opponent." Herman Kahn, *On Thermonuclear War* (Princeton University Press, 1960), 230.

[34] It would be a complication, at the least, that in a MAD-type deterrent standoff involving ASI, the superintelligence *itself* might be presumed to have agency. Even if with respect to "our" ASI, therefore, would *its* behavioral payoff matrices align completely with our own? And if not, what would happen? (What would *the ASI* want to accomplish in such circumstances, and would this be what we ourselves wished? Would it remain a tool in *our* MAD standoff with another power, or would *we* end up being tools in *its* relationship with our rival power's own ASI?) Particularly if the two ASIs involved in such a MAD world were both smarter than we are, it is presumably inherently hard to imagine how their relationship would develop. In any event, we are not yet at the point of anyone having and weaponizing some form of superintelligent capability.

Nor, for the reasons suggested earlier, is there any guarantee that we could get through the point of *someone* acquiring "first-mover" advantage in ASI without seeing its preemptive use against any "fast followers" before they acquired such a capability too. The United States did *not* choose nuclear preemption against the USSR during the brief years of 1945-49 when it enjoyed a nuclear monopoly, but who is to say what the ASI "first mover" of tomorrow might choose? (Nor is it clear what ASI "strategic preemption" would even mean in the first place. One of the problems of hypothesizing about superintelligence is that since it would by definition be more intelligent than we are, it is hard to know exactly what having it "on your side" in competition would allow you to do.)

[35] Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction (opened for signature April 10, 1972) (entered into force March 26, 1975), Art. I, https://treaties.unoda.org/t/bwc.

[36] Guy Roberts, "The Failure of the Biological Weapons Convention Protocol and a New Paradigm for Fighting the Threat of Biological Weapons," INSS *Occasional Paper* No. 49 (March 2003), ix, https://apps.dtic.mil/sti/tr/pdf/ADA435071.pdf.

[37] Hendrycks et al., "Superintelligence Strategy," 17-20.

[38] Kokotajlo et al, "AI 2027," 5-6.